Two stage adaptive cluster sampling

Two-Stage Adaptive Cluster Sampling

STEFANIA NADDEO *Università di Siena, Dipartimento di Metodi Quantitativi,*

NADDEO@UNISI.IT

Phone: 0577 232628
Fax: 0577 232626

CATERINA PISANI()) *Università di Siena, Dipartimento di Metodi Quantitativi*,

PISANI4@UNISI.IT

Phone: 0577 232750
Fax: 0577 232626

Abstract

Adaptive cluster sampling is usually applied when estimating the abundance of elusive, clustered biological populations. It is commonly supposed that all individuals in the selected area units are detected by the observer, but in many actual situations this assumption may be highly unrealistic and some individuals may be missed. This paper deals with the problem of handling imperfect detectability in adaptive cluster sampling by using a pure design-based approach. A two-stage adaptive procedure is proposed where the abundance in the selected units is estimated by replicated counts.

Key Words *adaptive cluster sampling* *imperfect detectability* *two-stage estimation* *replicated counts*

1. Introduction

Adaptive cluster sampling offers a suitable solution to the problem of estimating the abundance of rare, clustered populations. The design involves selecting an initial sample of area units and then adding neighbouring units whenever a given number of individuals is recorded within.

It is worth noting that adaptive cluster sampling is based on the assumption that every member of the population in the selected units is observed. In many real situations this assumption may be unrealistic, such as when dealing with barely detectable animals or elusive individuals, although the inference is usually made as if detectability were perfect.

The aim of this paper is to check the performance of adaptive cluster sampling in a realistic situation in which abundance in the selected units is estimated instead of being recorded without errors. In Section 2, available model-based suggestions concerning adjusting for imperfect detectability are described, while in Section 3 a two-stage adaptive cluster sampling design is proposed in a completely design-based setting. In the first stage, an initial sample of units is selected by means of simple random sampling without replacement, while the second stage involves estimating abundance within the sampled units by means of replicated encounter strategies. Accordingly, the total number of units included in the final sample depends on the values of the resulting estimates. The statistical properties of the derived estimator are then considered and subsequently, in Section 4 a simulation study is carried out to check the performance of the two-stage strategy. Finally, some remarks on the use of classical adaptive cluster sampling without perfect detectability are made in Section 5.

2. Preliminaries regarding Imperfect Detectability

Imperfect detectability is a problem frequently encountered in many surveys of natural and human populations, *i.e.* even if an area unit is included in the sample, all the individuals in the selected unit may not be detected by the observer. Some examples of imperfect detectability are aerial surveys of wild animals (Caughley, 1974, Caughley and Goddard, 1972, Routledge, 1981), vessel surveys of cetaceans, trawl surveys of fish, feasibility surveys of mining resources, surveys of artefacts in archaeological sites, surveys of homeless people (cfr. Thompson and Seber, 1994 and Seber and Thompson, 1994) and so on.

In conventional adaptive designs, imperfect detectability can influence the selection of the area units. If this is not taken into account, it may lead to an underestimate of population abundance. In order to handle imperfect detectability, Thompson and Seber

(1994) assume a detection probability for each individual of the population and modify the classical adaptive estimator by taking the detection probabilities into account. If the probabilities of detection are assumed to be known, the estimator of abundance turns out to be unbiased and its variance may be decomposed into two parts. It is at once apparent that one part of the variance depends on the detection probabilities while the other part depends on both the detection probabilities and the adaptive design adopted to select units (see Thompson and Seber, 1994, equations (2) and (10), p.713 and 719). In a realistic situation in which the detection probabilities are not known and must be estimated, the authors suppose that a consistent estimator of detectability may be obtained in a separate study. In this case, the abundance estimator may be proven to be approximately unbiased with an approximate variance which can be decomposed into three parts. In addition to the previous two sources of variability, the third term of the variance simply represents the increase of variability due to the estimation of the detection probabilities (see Thompson and Seber, 1994, equations (5) and (13), p.713 and 719).

As to the approach suggested by these authors, much research has been devoted to estimating detection probabilities both in a parametric and non-parametric setting, but the proposed solutions provide rather unsatisfactory results. In fact, parametric methods are very accurate if the model is properly selected but can show poor performance otherwise. On the contrary, the non-parametric methods give rise to robust estimates which often are not very accurate.

Hence, in this paper the issue of imperfect detectability is investigated by using a pure design-based approach, *i.e*. without assuming any model for the detection function. Particularly, the abundance in the units is estimated by means of plot sampling and the performance of the resulting two-stage adaptive estimator is evaluated on the basis of a simulation study.

3. Two-stage Adaptive Estimator

Consider a study region partitioned into N spatial units and denote by $T_1, T_2, \ldots, T_N$ the unit abundance. Let T be the whole abundance over the study area. An initial sample of n units is selected by simple random sampling without replacement. If

$\{1,2,\ldots,N\}$ denotes the set of indexes labelling the population units, then the initial sample may be viewed as a set of indexes $S_0 \subset \{1,2,\ldots,N\}$. Note that the abundance in each selected unit $l \in S_0$ is not observed but is instead estimated through an encounter sampling strategy. If the encounter procedure is independently replicated m_l times and the corresponding Horvitz-Thompson abundance estimate is subsequently computed, then the m_l replications give rise to m_l *iid* random variables with expectation T_l and variance σ_l^2 . Thus the sample mean of the m_l estimates, say $\hat{T}_l$, represents the realization of a random variable with expectation T_l and variance σ_l^2 / m_l. Moreover, $\hat{T}_l$ is asymptotically normal ($m_l \to \infty$) and an unbiased estimator of its variance is s_l^2 / m_l, where s_l^2 is the unbiased sample variance of the m_l estimates. Whenever $\hat{T}_l$ satisfies a given condition $\hat{T}_l \in C$ (*e.g.* $\hat{T}_l > 0$), additional units in the neighbourhood of the *l*-th unit are added to the sample. For each additional unit *k*, if $\hat{T}_k \in C$, the neighbouring units are also observed, and so on until a final sample is obtained. Note that the final sample is composed of clusters of units, each of which is formed by a boundary of units in which the estimate does not satisfy the condition (the so-called edge units) and by a network of units whose estimates satisfy condition *C*. Although $\hat{T}_l$s are quantified only for units included in the sample, it is mathematically convenient to define the vector $\hat{\mathbf{T}} = \left[\hat{T}_1, \hat{T}_2, \ldots, \hat{T}_N\right]^{\mathrm{T}}$. It is worth noting that since the estimation in each unit is performed by separate surveys, the component of $\hat{\mathbf{T}}$ are independent random variables.

As suggested by Thompson and Seber (1996), let us consider any unit not satisfying *C* as a network of size one, so that the population may be partitioned into networks. Let $U(\hat{\mathbf{T}})$ be the random partition of the population of units into networks, in such a way that, whenever $U(\hat{\mathbf{T}})$,

$$T = \sum_{j \in U(\hat{\mathbf{T}})} T_j^* = \sum_{l=1}^{N} T_l \quad ,$$

where T_j^* is the total abundance of the *j*-th network - that is, the sum of the abundance of the $n_j(\hat{\mathbf{T}})$ units belonging to the *j*-th network. Hence, the probability that the initial sample intersects the *j*-th network $\left(j \in U(\hat{\mathbf{T}})\right)$ turns out to be

$$\alpha_j(\hat{\mathbf{T}}) = 1 - \frac{\binom{N - n_j(\hat{\mathbf{T}})}{n}}{\binom{N}{n}}.$$

Henceforth, these probabilities will be denote by α_j for sake of simplicity.

Moreover, if $S(\hat{\mathbf{T}}) \subset U(\hat{\mathbf{T}})$ denotes the set of networks intersected by the initial sample, the two-stage estimator of the total is

$$\hat{\hat{T}} = \sum_{j \in S(\hat{T})} \frac{\hat{T}_j^*}{\alpha_j}, \tag{1}$$

where $\hat{T}_j^*$ is the estimator of the abundance of the *j*-th network, that is

$$\hat{T}_j^* = \sum_{i \in I_j(\hat{\mathbf{T}})} \hat{T}_i$$

and $I_j(\hat{\mathbf{T}})$ denotes the set of indexes labelling the units belonging to the *j*-th network.

As to the expectation of the two-stage estimator (1) it is at once apparent that

$$\mathrm{E}(\hat{\hat{T}}) = \mathrm{E}_{\hat{\mathbf{T}}}\left[\mathrm{E}_S(\hat{\hat{T}} \mid \hat{\mathbf{T}})\right]$$

and

$$\mathrm{E}_S(\hat{\hat{T}} \mid \hat{\mathbf{T}}) = \mathrm{E}_S\left(\sum_{j \in S(\hat{\mathbf{T}})} \frac{\hat{T}_j^*}{\alpha_j} \mid \hat{\mathbf{T}} \right) = \mathrm{E}_S\left(\sum_{j \in U(\hat{\mathbf{T}})} \frac{\hat{T}_j^*}{\alpha_j} Z_j \mid \hat{\mathbf{T}} \right) = \sum_{j \in U(\hat{\mathbf{T}})} \hat{T}_j^* ,$$

where E_S now denotes expectation with respect to the probability distribution induced by the design adopted to select S_0 and subsequently $S(\hat{\mathbf{T}})$, while Z_j is the indicator function which is equal to 1 if the initial sample intersects the j-th network and 0 otherwise. Thus, since $U(\hat{\mathbf{T}})$ constitutes a partition of $\{1,2,\ldots,N\}$, it is obvious that

$$\mathrm{E}(\hat{\hat{T}}) = \mathrm{E}_{\hat{\mathbf{T}}}\left[\sum_{j \in U(\hat{\mathbf{T}})} \hat{T}_j^* \right] = \mathrm{E}_{\hat{\mathbf{T}}}\left[\sum_{l=1}^{N} \hat{T}_l \right] = T .$$

Moreover, as to the variance of (1),

$$\mathrm{V}(\hat{\hat{T}}) = \mathrm{V}_{\hat{\mathbf{T}}}\left[\mathrm{E}_S(\hat{\hat{T}} \mid \hat{\mathbf{T}}) \right] + \mathrm{E}_{\hat{\mathbf{T}}}\left[\mathrm{V}_S(\hat{\hat{T}} \mid \hat{\mathbf{T}}) \right] =$$

$$= \mathrm{V}_{\hat{\mathbf{T}}}\left[\sum_{j \in U(\hat{\mathbf{T}})} \hat{T}_j^* \right] + \mathrm{E}_{\hat{\mathbf{T}}}\left[\sum_{j \in U(\hat{\mathbf{T}})} \frac{\hat{T}_j^{*2}(1-\alpha_j)}{\alpha_j} + 2 \sum_{j \in U(\hat{\mathbf{T}})} \sum_{h>j} \hat{T}_j^* \hat{T}_h^* \frac{(\alpha_{jh} - \alpha_j \alpha_h)}{\alpha_j \alpha_h} \right] =$$

$$= \sum_{l=1}^{N} \frac{\sigma_l^2}{m_l} + \mathrm{E}_{\hat{\mathbf{T}}}\left[\sum_{j \in U(\hat{\mathbf{T}})} \frac{\hat{T}_j^{*2}(1-\alpha_j)}{\alpha_j} + 2 \sum_{j \in U(\hat{\mathbf{T}})} \sum_{h>j} \hat{T}_j^* \hat{T}_h^* \frac{(\alpha_{jh} - \alpha_j \alpha_h)}{\alpha_j \alpha_h} \right], \qquad (2)$$

where, for sake of simplicity α_{jh} denotes

$$\alpha_{jh}(\hat{\mathbf{T}}) = 1 - \frac{\binom{N - n_j(\hat{\mathbf{T}})}{n} + \binom{N - n_h(\hat{\mathbf{T}})}{n} - \binom{N - n_j(\hat{\mathbf{T}}) - n_h(\hat{\mathbf{T}})}{n}}{\binom{N}{n}}$$

i.e. the probability that the initial sample intersects both the *j*-th and the *h*-th networks. Note that (2) differs from the variance of the classical adaptive estimator (Thompson and Seber, 1996). The first term depends on the estimation within all the units while the second term depends on both the selection of the initial sample and the estimated abundance in each unit. Moreover, (2) cannot be further developed straightforwardly since the involved quantities in the second term are random variables. However, an unbiased estimator of (2) may be straightforwardly obtained by

$$\hat{\mathrm{V}} = \sum_{j \in S(\hat{\mathbf{T}})} \frac{s_j^{*2}}{\alpha_j} + \sum_{j \in S(\hat{\mathbf{T}})} \frac{\hat{T}_j^{*2}(1-\alpha_j)}{\alpha_j^2} + 2 \sum_{j \in S(\hat{\mathbf{T}})} \sum_{h>j} \hat{T}_j^* \hat{T}_h^* \frac{(\alpha_{jh} - \alpha_j \alpha_h)}{\alpha_j \alpha_h \alpha_{jh}},$$

where

$$s_j^{*2} = \sum_{i \in I_j(\hat{\mathbf{T}})} \frac{s_i^2}{m_i} \quad .$$

Indeed

$$\sum_{j \in S(\hat{\mathbf{T}})} \frac{s_j^{*2}}{\alpha_j} \tag{3}$$

is an unbiased estimator of the first term of (2) since

$$\mathrm{E}\left(\sum_{j \in S(\hat{\mathbf{T}})} \frac{s_j^{*2}}{\alpha_j} \right) = \mathrm{E}_{\hat{\mathbf{T}}} \left[E_S \left(\sum_{j \in S(\hat{\mathbf{T}})} \frac{s_j^{*2}}{\alpha_j} \mid \hat{\mathbf{T}} \right) \right] = \mathrm{E}_{\hat{\mathbf{T}}} \left[E_S \left(\sum_{j \in U(\hat{\mathbf{T}})} \frac{s_j^{*2}}{\alpha_j} Z_j \mid \hat{\mathbf{T}} \right) \right] =$$

$$= \mathrm{E}_{\hat{\mathbf{T}}} \left(\sum_{j \in U(\hat{\mathbf{T}})} \sum_{i \in I_j(\hat{\mathbf{T}})} \frac{s_i^2}{m_i} \right) = \mathrm{E}_{\hat{\mathbf{T}}} \left(\sum_{i=1}^{N} \frac{s_i^2}{m_i} \right) = \sum_{i=1}^{N} \frac{\sigma_i^2}{m_i} .$$

Moreover

$$\sum_{j\in S(\hat{\mathbf{T}})} \frac{\hat{T}_j^{*2}(1-\alpha_j)}{\alpha_j^2} + 2\sum_{j\in S(\hat{\mathbf{T}})}\sum_{h>j} \hat{T}_j^{*}\hat{T}_h^{*}\frac{(\alpha_{jh}-\alpha_j\alpha_h)}{\alpha_j\alpha_h\alpha_{jh}} \tag{4}$$

is the Horvitz-Thompson estimator of the term between square brackets in the second term of (2), and as such it is unbiased with respect to the probability distribution induced by the design adopted to select S_0 and given the realization of $\hat{\mathbf{T}}$. Thus, (4) constitutes an unbiased estimator of the second term of (2).

4. Some Monte Carlo Simulations

In order to check the performance of the two-stage procedure proposed in the previous section, the artificial population of N=400 squared units of size one described by Thompson (1992, p.285) was considered and the individuals in each unit were allocated in the nodes of a regular grid. Then, 10,000 initial samples of size n=5(5)20 were selected by simple random sampling without replacement. For each selected sample, both the classical adaptive cluster strategy as well as the two-stage strategy were performed. As to the two-stage strategy, the abundance within the selected units was estimated by a plot sampling procedure performed using m_j=10(10)30 circular plots with radius r=0.06.

The empirical variances (EV) of the two-stage abundance estimator were computed together with the empirical expectations of the effective surveyed surface (ESS) - that is, the surface of the circular plot multiplied by the overall number of plots allocated in the selected units - arising from the adaptive procedure on the basis of the 10,000 samples. Moreover, the exact variance (V) and expected sample sizes (SS) of the classical adaptive estimator were theoretically determined. It is at once apparent that in the classical adaptive procedure the expected sample size corresponds to the expected effective surveyed surface, since the surface of each selected unit is completely investigated.

Note that an empirical evaluation of the variability due to the estimation of the totals within the units is given by the difference between the empirical variance of the two-stage estimator and the exact variance of the classical estimator, which is EV-V.
The results of the simulation are reported in Table 1, where in order to emphasize the performance of the two-stage procedure with respect to its classical counterpart, the relative increase in variability (RIV) due to estimation within the sampled units is reported together with the average relative decrease of the effective surveyed surface (RDS).

Table 1

5. Concluding Remarks

From the previous results it is at once apparent that the variance of the two-stage estimator is dramatically higher than the variance of the classical estimator and the increase in variability cannot be explained even by taking into account the decrease in the effective surveyed surface. Obviously, as the initial sample size increases, the variance of the classical adaptive estimator considerably decreases, while both the initial sample size and the number of plots allocated for each selected unit affect the values of the variance of the two-stage estimator.
On the basis of the simulation results, it is worth noting that the average increase of variability due to the estimation of abundance (RIV) is about 25 when 10 plots are allocated to each selected unit and falls to 12.11 and 7.20 when 20 or 30 plots are considered. Moreover, from an analysis of the last column of Table 1, the reduction of the effective surveyed surface is found to be quite stable with respect to the initial sample size and varies from about 0.94 for 10 plots to about 0.80 for 30 plots.
It should be noted that the increase in variability of the two-stage adaptive estimator may also be explained by taking into account that, since the abundance in each selected unit is estimated, even if a network is intercepted by the initial sample, it is possible that some of the units belonging to the network are not included in the final sample.

Thus, in accordance with these considerations, when adopting adaptive cluster sampling without perfect detectability, the increase of the variance due to the estimation within the selected units should not be neglected, as this involves unreliably evaluating the precision of the resulting estimates. Accordingly, since the assumption of perfect detectability may be highly unrealistic when dealing with elusive populations, the estimation of the overall variability, including that due to the estimation of the abundance in the selected units, would appear imperative in order to avoid dangerous underestimation of the sampling variance and subsequent excessive confidence in the effectiveness of the sampling strategy.

These results seem to be in great contrast with those reported in the paper by Jensen (1996) where a simulation study on the performance of two-stage line-transect sampling is described. The simulation results suggest that the variance due to the estimation procedure within the selected units is insignificant when compared to the overall variance if the number of selected units is very small compared to the number of units partitioning the study area. Hence, the author suggests estimating the total variance "using only the replications among transects".

It is worth noting that the apparent contrast between Jensen's and our remarks may be explained by considering that his results depend heavily on the detection function adopted in the simulation. Different results might have been obtained with less favourable detection functions, *e.g.* when adopting functions in which visibility markedly falls as distance increases.

References

Caugley, G. (1974) Bias in Aerial Survey, *Journal of Wildlife Management*, **38**, 921-933.

Caugley, C. and Goddard, J. (1972) Improving the Estimate from Inaccurate Censuses, *Journal of Wildlife Management*, **36**, 135-140.

Jensen, A.L. (1996) Subsampling with Line Transects for Estimation of Animal Abundance, *Environmetrics* ,**7**, 283-289.

Routledge, R.D. (1981) The Unreliability of Population Estimates from Repeated, Incomplete Aerial Surveys, *Journal of Wildlife Management*, **45**, 997-1000.

Seber, G.A.F., Thompson, S.K. (1994) Environmental Adaptive Sampling, *Handbook of Statistics*, Vol. 12, 201-220, G.P. Patil and C.R. Rao eds.

Thompson, S.K. (1992) *Sampling*, John Wiley and Sons.

Thompson, S.K., Seber, G.A.F. (1994) Detectability in Conventional and Adaptive Sampling, *Biometrics*, **50**, 712-724.

Thompson, S.K., Seber, G.A.F. (1996) *Adaptive Sampling*, John Wiley and Sons.

Table 1: Empirical comparison of the two-stage adaptive procedure with respect to the classical adaptive procedure.

n	number of plots	Two-stage Adaptive Procedure		Classical Adaptive Procedure		RIV	RDS
		EV	ESS	V	SS		
5	10	2,644,895.33	0.58	131,257.92	9.39	19.15	0.94
	20	1,408,257.42	1.18	131,257.92	9.39	9.73	0.87
	30	1,090,103.64	1.83	131,257.92	9.39	7.30	0.81
10	10	1,799,755.19	1.15	61,848.27	18.26	28.10	0.94
	20	817,594.05	2.36	61,848.27	18.26	12.22	0.87
	30	463,358.77	3.63	61,848.27	18.26	6.49	0.80
15	10	1,056,190.22	1.73	38,805.13	26.67	26.22	0.94
	20	530,972.40	3.55	38,805.13	26.67	12.68	0.87
	30	308,223.99	5.44	38,805.13	26.67	6.94	0.80
20	10	756,059.64	2.31	27,355.20	34.66	26.64	0.93
	20	405,427.28	4.72	27,355.20	34.66	13.82	0.86
	30	248,325.45	7.24	27,355.20	34.66	8.08	0.79